Inversio-ongelmasta ja parametrien estimoinnista

LuK-tutkielma
Iida Välipirtti
2503398
Matemaattisten tieteiden tutkinto-ohjelma
Oulun yliopisto
Kevät 2019

Sisältö

Johdanto

Tässä tutkielmassa tutustutaan inversio-ongelmiin, ja tapoihin ratkaista niitä. Aluksi esitellään inversio-ongelma ja avataan syitä sille, miksi se on hankala ratkaista. Tämän jälkeen tutustutaan tilastollisiin menetelmiin ratkaista inversio-ongelmia ja arvioida mallien oikeutta. Lineaarinen regressio esitellään parametrien estimointiongelmana, ja pienimmän neliösumman ratkaisu johdetaan siitä. Suurimman uskottavuuden estimointi esitellään, kuten myös tuntemattoman keskihajonnan ongelmat. Tutkielmassa on käytetty päälähteenä teosta [1].

1 Inversio-ongelma

1.1 Johdattelua aiheeseen

Olkoon d havainnoista kerätty aineisto, jonka pohjalta halutaan muodostaa tietyillä parametreilla malli m. Olkoon G kuvaus

$$G(m) = d. \tag{1}$$

Käytännön tilanteissa d voi olla esimerkiksi funktio ajan tai paikan suhteen, tai joukko diskreettejä havaintoja. On myös muistettava, että havainnoissa on käytännössä aina mukana virhettä. Tällöin voidaan ajatella aineiston koostuvan sekä virheettömistä havainnoista että virhekomponentista η,

$$d = G(m_{tosi}) + \eta \tag{2}$$
$$= d_{tosi} + \eta \tag{3}$$

missä d_{tosi} toteuttaa kaavan (1), kun m on virheetön malli m_{tosi}, ja malli on oikein laadittu. Löydetty malli ei kuitenkaan ole ainoa laatuaan, vaan yleensä voidaan muodostaa äärettömästi malleja, jotka sopivat aineistoon d_{tosi}. Kun m ja d ovat funktioita, funktiota G kutsutaan operaattoriksi. Operaattori G voi olla differentiaaliyhtälö, osittaisdifferentiaaliyhtälö, lineaarinen yhtälöryhmä tai epälineaarinen yhtälöryhmä.

Matemaatikoiden ja muiden tieteilijöiden mallinnukseen liittyvä termistö eroaa toisistaan. Sovelletun matematiikan parissa yhtälöä (1) kutsutaan *matemaattiseksi malliksi*, ja muuttujaa m *parametriksi*. Soveltavilla aloilla, kuten muissa luonnontieteissä ja insinööritieteissä, operaattoria G kutsutaan *suoraksi operaattoriksi* ja muuttujaa m *malliksi*. Tässä tutkielmassa muuttujaa m kutsutaan *malliksi* ja yhtälöä (1) *matemaattiseksi malliksi*.

Usein kiinnostuksenkohteena on tapaus, jossa halutaan tunnetun mallin m avulla ratkaista d. Tämänkaltaista tilannetta kutsutaan suoraksi ongelmaksi. Tässä tutkielmassa keskitytään sen sijaan inversio-ongelmaan, eli käänteiseen ongelmaan, jossa tunnetun aineiston d avulla halutaan ratkaista m.

1.2 Inversio-ongelmien luokittelua

Mallin määrittämiseen tarvitaan usein äärellinen määrä parametreja n. Tässä tapauksessa mallin parametreja voidaan kuvata k-alkioisena vektorina m. Jos tunnetaan äärellinen määrä havaintopisteitä, aineistoa voidaan kuvata l-ulotteisena vektorina d. Tämänkaltaisia ongelmia kutsutaan diskreeteiksi

inversio-ongelmiksi tai parametrien estimointiongelmiksi. Yleinen parametrien estimointiongelma voidaan kirjoittaa yhtälöparina

$$G(m) = d. \tag{4}$$

Tapauksissa, joissa malli ja aineisto ovat aika-avaruuden funktioita, mallin m estimointia aineistosta d kutsutaan jatkuvaksi inversio-ongelmaksi. Ongelmia voidaan luokitella myös siihen liittyvien parametrien määrän perusteella: jos parametreja on vähän, puhutaan parametrien estimointiongelmista, kun taas useiden parametrien tilanteessa ongelmaa kutsutaan inversio-ongelmaksi.

Yleisesti hyödyllisiksi matemaattisiksi malleiksi on todettu lineaariset systeemit, jotka toteuttavat superpositioperiaatteen

$$G(m_1 + m_2) = G(m_1) + G(m_2) \tag{5}$$

sekä skaalauksen

$$G(\alpha m) = \alpha G(m). \tag{6}$$

Diskreetin lineaarisen inversio-ongelman tapauksessa yhtälö (4) voidaan kirjoittaa lineaarisena yhtälöryhmänä

$$G(m) = Gm = d. \tag{7}$$

Jatkuvan lineaarisen inversio-ongelman tapauksessa, operaattori G voidaan kuvata lineaarisena integrointioperaattorina, jossa (1) on muotoa

$$\int_a^b g(s,x)m(x)dx = d(s) \tag{8}$$

ja funktiota $g(s,x)$ kutsutaan ytimeksi. Kaava (8) on lineaarinen, sillä

$$\begin{aligned} &\int_a^b g(s,x)(m_1(x) + m_2(x))dx \\ &= \int_a^b g(s,x)m_1(x)dx + \int_a^b g(s,x)m_2(x)dx \end{aligned} \tag{9}$$

ja

$$\int_a^b g(s,x)\alpha m(x)dx = \alpha \int_a^b g(s,x)m(x)dx. \tag{10}$$

Muotoa (8) olevia yhtälöitä, joissa $m(x)$ on tuntematon, kutsutaan ensimmäisen lajin Fredholmin integraaliyhtälöiksi. Kyseisistä integraaliyhtälöistä

on muodostettu lukuisia inversio-ongelmia, sillä niistä on vaikea saada hyödyllisiä ratkaisuja suorilla menetelmillä.

Kaavan (8) ydin voidaan kirjoittaa riippumaan eksplisiittisesti muuttujasta $s - x$, jolloin saadaan konvoluutioyhtälö

$$\int_{-\infty}^{\infty} g(s-x)m(x)dx = d(s). \tag{11}$$

Suoraa ongelmaa, jossa $d(s)$ ratkaistaan $m(x)$ avulla, ja joka on muotoa (11) kutsutaan konvoluutioksi. Vastaavaa inversio-ongelmaa, jossa $m(x)$ ratkaistaan $d(s)$ avulla, kutsutaan dekonvoluutioksi.

Eräs ensimmäisen lajin Fredholmin integrointiyhtälöön pohjautuva ongelma on Fourier-muunnoksen

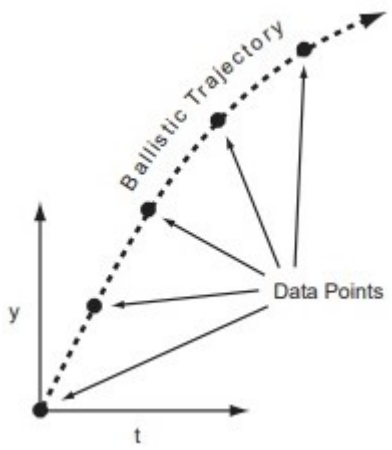
$$\Phi(f) = \int_{-\infty}^{\infty} \exp(-i2\pi fx)\phi(x)dx, \tag{12}$$

kääntäminen, jotta saataisiin ratkaistua $\phi(x)$. Fourier-muunnoksien ja niiden käänteisfunktioiden ratkaisemista varten on olemassa taulukoita ja analyyttisiä menetelmiä, mutta kiinnostavaa olisi saada esimerkiksi numeerinen estimaatti muunnokselle $\phi(x)$, vaikka ei tiedettäisi sen analyyttistä käänteisfunktiota.

1.3 Esimerkkejä

Käsitellään kaksi esimerkkiä; ensimmäinen on parametrien estimoinnista, ja toinen inversio-ongelmasta.

Esimerkki 1.1. Yksinkertainen parametrien estimointiongelma on sovittaa parametrien määrittämä funktio aineistoon. Tapauksia, joissa tämä sovitus onnistuu, kutsutaan lineaariseksi regressioksi. Muinainen esimerkki lineaarisesta regressiosta on lentoradan luonnostelu. Tällöin aineisto y koostuu korkeuden mittauksista ajanhetkellä t.



Kuva 1. Parabolisen liikeradan ongelma.

Tavoitteena on löytää malli m, joka sisältää alkukorkeuden k_1, alkunopeuden pystysuunnassa k_2 sekä putoamiskiihtyvyyden k_3. Tämän tilanteen matemaattinen malli on neliöllinen funktio (t, y) -tasolla

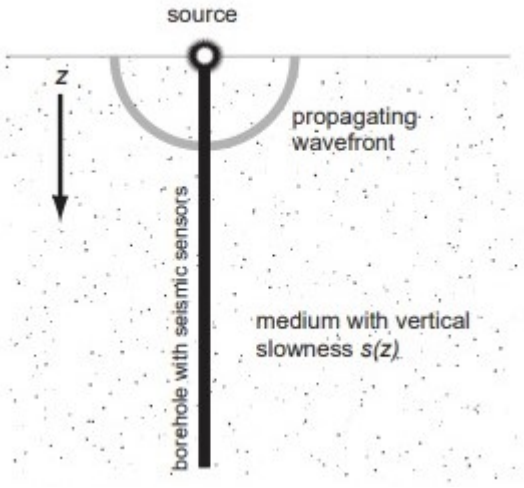
$$y(t) = k_1 + k_2 t - (1/2)k_3 t^2. \tag{13}$$

Aineisto sisältää k havaintoa y_i tutkittavan kappaleen korkeudesta ajanhetkillä t_i. Olettaen, että t_i on mitattu tarkasti, ja käyttäen kaavaa (13), saamme yhtälöparin, joka sopii aineistoomme y_i malliparametreilla k_j:

$$\begin{bmatrix} 1 & t_1 & -\frac{1}{2}t_1^2 \\ 1 & t_2 & -\frac{1}{2}t_2^2 \\ 1 & t_3 & -\frac{1}{2}t_3^2 \\ \vdots & \vdots & \vdots \\ 1 & t_k & -\frac{1}{2}t_k^2 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_k \end{bmatrix}. \tag{14}$$

Vaikka (13) on neliöllinen, kolmen parametrin k_i yhtälöt ovat lineaarisia, joten mallin m ratkaiseminen on lineaarinen parametrien estimointiongelma.

Esimerkki 1.2. Geofysiikassa ollaan kiinnostuneita seismologiasta, joka tutkii maanjäristyksiä ja muita maan läpi kulkevia aaltoliikkeitä. Siihen kuuluvaa pystysuuntaista seismistä profilointia tarkastellessa halutaan tietää porausreiän ympärillä olevan materiaalin seisminen nopeus. Ongelmaa varten suoritetaan mittauksia.



Kuva 2. Pystysuuntainen seisminen profilointiongelma.

Ongelma on epälineaarinen, jos se ratkaistaan nopeusparametrien pohjalta. Ongelma voidaan linearisoida parametrisoimalla hitaus s nopeuden v sijaan. Havaittu matkaan kulunut aika syvyydessä z voidaan kuvata pystysuuntaisen hitaden s määrätyllä integraalilla, kun valitaan integrointiväliksi maanpinnasta syvyyteen z,

$$t(z) = \int_0^z s(\xi)d\xi \tag{15}$$

$$= \int_0^\infty s(\xi)H(z-\xi)d\xi. \tag{16}$$

Kaavassa (16) esiintyvä ydinfunktio H on Heavisiden funktio

$$H(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0. \end{cases} \tag{17}$$

Teoriassa, (16) voidaan ratkaista melko helposti, sillä analyysin peruslauseen perusteella

$$t'(z) = s(z). \tag{18}$$

Todellisissa tapauksissa aineistossa voi olla mukana häiriöitä, jolloin havaintojen derivointi voi tuottaa erittäin häiriöisen ratkaisun.

1.4 Miksi inversio-ongelmat ovat hankalia?

Kuten aiemmin jo mainittiin, yleensä on mahdollista määrittää lukuisia malleja, jotka sopivat riittävän hyvin mallintamaan tiettyä aineistoa. On siis tärkeää luonnehtia saatua ratkaisua, ja selvittää kuinka uskottava ja aineistoon

sopiva se on. Kannattaa siis tarkastella kolmea ehtoa: ratkaisun olemassaoloa, sen yksikäsitteisyyttä sekä ratkaisumenetelmän stabiilisuutta.

1. Olemassaolo: On mahdollista ettei ole olemassa täydellistä, tai edes riittävän lähellä olevaa, mallia kuvaamaan aineistoa. Syitä tähän voi olla esimerkiksi aineistossa esiintyvät häiriöt.

2. Yksikäsitteisyys: Vaikka tarpeeksi tarkka malli saataisiin muodostettua, on mahdollista havaita, ettei se ole yksikäsitteinen. Voi siis olla olemassa sekä m_1 että m_2, jotka toteuttavat yhtälön $G(m_i) = d$, kun $i = 1, 2$ ja $m_1 \neq m_2$. Ongelman ei-yksikäsitteisten ratkaisujen haittapuolena on se, että estimoidut mallit voivat olla liian sovitettuja tai harhaisia verrattuna oikeaan malliin. Harhan tarkastelu onkin tärkeää, kun tutkitaan mallien ja todellisuuden mahdollista vastaavuutta.

3. Stabiilisuus: Inversio-ongelman ratkaisun laskeminen on usein epästabiilia, sillä pienikin muutos mittauksissa voi johtaa estimoidun mallin merkittävään muuttumiseen. Tämän vuoksi inversio-ongelma on häiriöaltis.

Esimerkki 1.3. Tarkastellaan ensimmäisen lajin Fredholmin integraaliyhtälöä

$$\int_0^1 g(s,x)m(x)dx = y(s). \tag{19}$$

Triviaalissa tapauksessa $g(s,x) = 1$, jolloin integraaliyhtälö sievenee muotoon

$$\int_0^1 m(x)dx = y(s). \tag{20}$$

Koska yhtälön (20) vasen puoli ei riipu muuttujasta s, systeemillä ei ole ratkaisua ellei $y(s)$ ole vakio. Saamme siis kaksi tapausta:

1. $y(s)$ ei ole vakio.

   On olemassa äärettömästi aineistojoukkoja, jotka eivät ole vakioita ja siksi niille ei ole ratkaisua. Kyseessä on siis ristiriita ratkaisun olemassaolon kanssa.

2. $y(s)$ on vakio.

   Koska on olemassa äärettömästi funktioita, joiden integraali tuottaa saman vakion, niin $y(s)$ ei ole ainutlaatuinen. Kyseessä on siis ristiriita ainutlaatuisuuden kanssa.

2 Lineaarinen regressio

2.1 Johdatus lineaariseen regressioon

Jos tavoitteena on muodostaa parametrisoitu käyrä, joka likimäärin sopii aineistoon, kyseessä on regressio-ongelma. Kun regressiomalli on lineaarinen, käsittelemämme tapaus on nimeltään lineaarinen regressio-ongelma.

Käsitellään diskreettiä lineaarista inversio-ongelmaa. Olkoon d aineistovektori, joka sisältää k havaintoa, ja olkoon vektori m mallivektori, joka sisältää l parametria, jotka haluamme selvittää. Kuten aiemmin on jo todettu, inversio-ongelma voidaan kirjoittaa lineaarisena yhtälöryhmänä

$$Gm = d. \tag{21}$$

Residuaaliin johdattelussa tarvitaan matriisien käsitteitä. Määritellään matriisin sarakeaste.

Määritelmä 2.1. Olkoon matriisi A $a \times b$ –matriisi. Matriisin sarakeaste kuvaa sen lineaarisesti riippumattomien sarakevektoreiden määrää. Astetta merkitään $rank(A)$. Jos $rank(A) = b$, niin matriisilla A on täysi sarakeaste.

Oletetaan, että matriisilla G on täysi sarakeaste. Tällaisissa tilanteissa on yleistä, ettei ratkaisuksi saatu malli m toteuta täysin yhtälöä (21). Hyvä approksimaatio saadaan kuitenkin etsimällä sellainen malli m, joka minimoi sopimattomuutta varsinaisen aineiston ja arvion Gm välillä. Residuaalivektoria, jonka alkioita kutsutaan residuaaliksi eli jäännösvirheeksi, merkitään

$$r = d - Gm. \tag{22}$$

Eräs usein käytetty epäsopivuuden mitta on residuaalien normi. Mallia, joka mimimoi tämän normin, kutsutaan pienimmän neliösumman ratkaisuksi. Sen on todettu olevan tilastollisesti uskottavin ratkaisu olettaen, että aineiston sisältämät virheet ovat normaalijakautuneita. Pienimmän neliösumman ratkaisu on yleisesti muotoa

$$m_{L_2} = (G^T G)^{-1} G^T d. \tag{23}$$

Yleinen lineaarisen regression ongelma on löytää suoralle $y = k_1 + k_2 x$ sellaiset parametrit k_1 ja k_2, joilla suora sopii mahdollisimman hyvin kuvaamaan havaintoja. Nyt yhtälöryhmä on muotoa

$$Gm = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_k \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{bmatrix} = d. \tag{24}$$

Yhtälöä (23) käyttämällä pienimmän neliosumman ratkaisu antaa tulokseksi

$$\begin{aligned} m_{L_2} &= (G^T G)^{-1} G^T d \\ &= \left(\begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_k \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_k \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_k \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{bmatrix} \\ &= \begin{bmatrix} k & \sum_{i=1}^k x_i \\ \sum_{i=1}^k x_i & \sum_{i=1}^k x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^k d_i \\ \sum_{i=1}^k x_i d_i \end{bmatrix} \\ &= \frac{1}{k \sum_{i=1}^k x_i^2 - \left(\sum_{i=1}^k x_i \right)^2} \begin{bmatrix} \sum_{i=1}^k x_i^2 & -\sum_{i=1}^k x_i \\ -\sum_{i=1}^k x_i & k \end{bmatrix} \begin{bmatrix} \sum_{i=1}^k d_i \\ \sum_{i=1}^k x_i d_i \end{bmatrix}. \end{aligned} \tag{25}$$

2.2 Pienimmän neliösumman tilastollisia näkökulmia

Jos havaintopisteet ovat epätarkkoja mittauksia, joissa on mukana satunnaisvirheitä, niin kuinka voidaan löytää tilastollisesta näkökulmasta paras ratkaisu? Suurimman uskottavuuden estimointi voisi olla eräs vaihtoehto. Sitä käytettäessä täytyy tietää havaintoiden avulla aineiston tilastollisia ominaisuuksia, ja tuntea suoran ongelman matemaattinen malli. Tehtävänä on selvittää sellainen malli, josta kyseiset havainnot ovat uskottavimmin peräisin.

Suurimman uskottavuuden estimointia voidaan soveltaa mihin tahansa estimointiongelmaan, jossa yhteistiheysfunktio

$$P(X \leq a \text{ ja } Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x,y) dy dx \tag{26}$$

voidaan määrittää havaintojen pohjalta. Oletetaan, että havainnot ovat keskenään riippumattomia, jolloin voidaan käyttää yhteistiheysfunktion tulomuotoa

$$f(x,y) = f_X(x) f_Y(y). \tag{27}$$

Mallin m perusteella ollaan muodostettu tiheysfunktio $f_i(d_i|m)$ jokaiselle havainnolle i. Nämä tiheysfunktiot vaihtelevat riippuen mallista m. Riippumattomien havaintojen vektorin d yhteistiheys on

$$f(d|m) = f_1(d_1|m) \cdot f_2(d_2|m) \cdots f_k(d_k|m). \tag{28}$$

Havaintoaineiston todennäköisyys mallille m, voidaan laskea integroimalla tiheyttä $f(d|m)$ annetun välin yli. Käytännön tilanteissa aineistovektoria

mitataan, ja toivotaan löydettävän paras versio sopivasta mallista suurimman uskottavuuden kannalta. Tässä tapauksessa d on kiinnitetty aineistojoukko, ja m on estimoitavat parametrit sisältävä vektori. Uskottavuusfunktio on

$$L(m|d) = f(d|m) \tag{29}$$
$$= f_1(d_1|m) \cdot f_2(d_2|m) \cdots f_k(d_m|m). \tag{30}$$

Yhtälön (30) tulos on useille malleille erittäin lähellä nollaa. Tällöin mallin on erittäin epätodennäköistä tuottaa havaittujen arvojen joukkoa. Tietyille malleille uskottavuus on selvästi suurempi, ja täten nämä mallit tuottavat todennäköisemmin juuri niitä arvoja, joita ollaan havaittu. Kyseiset mallit ovat suurimman uskottavuuden menetelmän mukaisia. Suurimman uskottavuuden menetelmän mukaan on siis valittava malli m, joka maksimoi uskottavuusfunktion (30) arvon. Mielenkiintoista on huomata, että jos käsitellään diskreettejä lineaarisia inversio-ongelmia, joiden havainnoissa esiintyvät virheet ovat normaalijakautuneita, niin niiden suurimman uskottavuuden menetelmän mukainen ratkaisu on myöskin pienimmän neliösumman menetelmän ratkaisu.

Jäännösvirheiden neliösumma tarjoaa hyödyllistä tilastollista tietoa pienimmän neliösumman menetelmällä saadun malliestimaatin laadusta. Khiin neliö -tunnusluku

$$\chi^2_{hav} = \sum_{i=1}^{k} (d_i - (Gm_{L_2})_i)^2/\sigma_i^2 \tag{31}$$

on eräs esimerkki tästä. Kaavassa (31) esiintyvä σ^2 kuvaa varianssia. Koska χ^2_{hav} riippuu satunnaisista mittausvirheistä aineistossa d, se on satunnaismuuttuja.

Khiin neliö -testi antaa tilastollista arviota oletuksista, joita käytettiin pienimmän neliösumman ratkaisemiseen. Kyseisessä testissä lasketaan arvo tunnusluvulle χ^2_{hav}, ja verrataan sitä teoreettiseen χ^2 – jakaumaan. Todennäköisyys sille, että teoreettinen arvo on yhtä suuri tai suurempi kuin havaittu arvo on

$$p = \int_{\chi^2_{hav}}^{\infty} f_{\chi^2}(x)dx. \tag{32}$$

Tätä kutsutaan p-arvoksi. Kun aineistossa esiintyvät virheet ovat riippumattomia sekä normaalijakautuneita, ja matemaattinen malli on oikein, p-arvo on tasaisesti jakautunut nollan ja yhden välille. Käytännön tilanteissa ne p-arvot, jotka ovat erittäin lähellä jompaakumpaa päätepistettä, viittaavat siihen, että jokin oletuksista on väärin.

On olemassa kolme vaihtoehtoa.

1. P-arvo ei ole erityisen suuri eikä pieni.

   Tässä tapauksessa pienimmän neliösumman ratkaisu on hyväksyttävissä ja tilastolliset oletukset virheiden laadusta ovat pitäviä. Käytännössä p-arvon ei tarvitse olla kovinkaan suuri, jotta ratkaisu tulee hyväksytyksi, sillä aidosti väärät mallit antavat erittäin pieniä (jopa luokkaa 10^{-12} olevia) arvoja normaalijakauman lyhyen hännän vuoksi.

   P-arvo on tasajakautunut, joten kun matemaattinen malli on oikein ja oletukset ovat päteviä, on virheellistä olettaa mitään p-arvon hyvyydestä. Esimerkiksi p-arvon 0,7 ja 0,2 välillä ei ole käytännössä mitään eroa.

2. P-arvo on hyvin pieni.

   Malli ja aineisto eivät vastaa tällöin toisiaan, johon voi olla useita syitä.

   (a) Aineisto kuvaa erittäin epätodennäköistä tilannetta.

   (b) Matemaattinen malli $Gm = d$ on väärin. Yleensä syynä on tilanteeseen sopimaton malli.

   (c) Aineiston virheet on aliarvioitu tai ne eivät ole normaalijakautuneita.

3. P-arvo on erittäin suuri eli lähes yksi. Nyt malli ja aineisto vastaavat lähes täydellisesti toisiaan, joten täytyy selvittää, onko virheitä yliarvioitu. Toinen erittäin harvinainen mahdollisuus on se, että aineistoa on muokattu vastaamaan jotain tiettyä mallia.

Joissain tapauksissa kannattaa Khiin neliön havaintojen χ^2_{hav} lisäksi tarkastella myös residuaalien ja mallin vastaavuutta mallin oikeuden tunnistamiseksi. Residuaalien tulisi olla normaalijakautuneita sekä keskihajonnan tulisi olla yksi. Residuaalien ei myöskään tulisi noudattaa mitään selkeää kaavaa. Esimerkiksi suoran lineaarisessa regressiossa residuaalit saattavat olla negatiisia kun riippumaton muuttuja x saa pieniä tai suuria arvoja, ja positiivia kun x saa puolivälissä olevia arvoja, jolloin regressiomallin voi todeta tarvitsevan esimerkiksi neliöllistä termiä.

2.3 Tuntemattomat mittauskeskihajonnat

Oletetaan, ettei tiedetä mittausvirheiden keskihajontoja ennen havaintoihin tutustumista. Oletetaan myös, että mittausvirheet ovat riippumattomia sekä normaalijakautuneita odotusarvolla 0 ja keskihajonnalla σ. Tällöin voimme suorittaa lineaarisen regression sekä estimoida keskihajontaa σ residuaalien avulla.

Aluksi on löydettävä pienimmän neliösumman ratkaisu ongelmalle $Gm = d$. Olkoon

$$r = d - Gm_{L_2}. \tag{33}$$

Residuaaleihin perustuvan keskihajonnan estimointiin tarvitaan kaavaa

$$s = \sqrt{\frac{1}{k-l}\sum_{i=1}^{k} r_i^2}. \tag{34}$$

Kun keskihajonta täytyy estimoida, tilastollinen tarkkuus heikkenee verrattuna siihen, että käytettäisi tarkkaa arvoa.

Tilanteessa, jossa aineistokeskihajonta σ tunnetaan, mallivirheet

$$m_i' = \frac{m_i - m_{tosi_i}}{\sigma} \tag{35}$$

noudattavat standardinormaalijakaumaa. Kun keskihajonnan σ tilalla käytetään sen kaavasta (34) saatua estimaattia s, mallivirheet

$$m_i' = \frac{m_i - m_{tosi_i}}{s} \tag{36}$$

noudattavat Studentin t-jakaumaa vapausasteella $k - l$. Mitä suurempi vapausaste on, sitä parempi estimaatti s saadaan.

Jos keskihajonta ei ole tunnettu, khiin neliö -testiä ei voida käyttää, sillä se vaatisi aineistovirheiden olevan normaalijakautuneita, ja keskihajonnan olevan tunnettu. Jos todelliset residuaalit olisivat liian suuria suhteessa keskihajontaan σ_i, niin χ^2 olisi suuri, ja jouduttaisiin hylkäämään lineaarisen regression sopivuus tilanteeseen liian pienen p-arvon vuoksi. Sijoittamalla keskihajonnan estimaatti (34) khiin neliö-tunnuslukuun (31), huomataan, että $\chi^2_{hav} = k - l$. Tällainen malli toteuttaa siis aina khiin neliö -testin.

Lähdeluettelo

[1] Aster R.C, Borchers B, and Thurber C.H: *Parameter Estimation and Inverse Problems*. Elsevier Academic Press, MA, 2005.